Building High-fidelity Human Body Models from User-generated Data

Zongyi Xu, Wei Chang, Yindi Zhu, Le Dong, Huiyu Zhou, Qianni Zhang

***Abstract*—We propose a key point-based approach, refers to as *KPhub-PC*, to estimate high-fidelity human body models from low-quality point clouds acquired with an affordable 3D scanner and a variation *KPhub-I* that can achieve the same purpose based on low-resolution single images taken by smartphones. In KPhub-PC, a sparse set of key points is annotated to guide the deformation of a parametric 3D human body model SMPL and then a high-fidelity human body model that can explain the target point cloud is built. Besides building 3D human body models from point clouds, KPhub-I is designed to estimate accurate 3D human body models from single 2D images. The SMPL model is fitted to 2D joints and the boundary of the human body which are detected using CNN based methods automatically. Considering that people are in stable poses most of the time, a stable pose prior is defined from CMU motion capture dataset for further improving accuracy. Extensive experiments demonstrate that in both types of user-generated data, the proposed approaches can build believable and animatable human body models robustly. Our approach outperforms the state-of-the-arts in the accuracy of both human body shape and pose estimation.**

***Index Terms*—Human body modelling, User-generated data, Point clouds, Single image, Virtual dressing.**

I. Introduction

NOwadays, to create the 3D avatar of a customer, current virtual dressing applications such as Metail [1], DressingRoom by Gap [2] or Biometrics [3] either rely on customers' manual entry of some body measurements which give limited details of body shape and unrealistic avatars, or scanning the customers' body shapes in a professional setup, like TEN 24 [4] or the 3D stereo capture system in Max Planck Institute [5], which is inaccessible to general users.

This work is supported in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant NO. KJQN201900628, and in part by Natural Science Foundation of Chongqing under No. E021D2019034. We would like to thank the Innovation Team on Digital Fashion and Spatial Visual Communication in School of Art & Design, Beijing Institute of Fashion Technology, for providing the original images of the IM-hub dataset of various participants and collecting the body physical measurements. (Corresponding author: Qianni Zhang.)

Zongyi Xu is with the Key Laboratory of Data Engineering and Visual Computing, Chongqing University of Posts and Telecommunications. (email: xuzy@cqupt.edu.cn)

Qianni Zhang is with the Department of Electrical Engineering and Computer Science, Queen Mary Univeristy of London, London, UK, E1 4NS. (email: qianni.zhang@qmul.ac.uk)

Wei Chang and Yindi Zhu are with Beijing Institute of Fashion Technology.

Le Dong is with University of Electronic Science and Technology of China.

Huiyu Zhou is with University of Leicester, United Kingdom.

[1]https://metail.com/

[2]http://www.dailymail.co.uk/sciencetech/article-4089338/Gap-reveals-new-app-lets-virtually-try-clothes-home.html

[3]http://bodymetrics.com/

[4]http://ten24.info/3d-scanning/

[5]https://ps.is.tuebingen.mpg.de/pages/3dcapture

With the appearance of commodity sensors, many capturing systems are proposed [1], [2], [3] to reconstruct the full body. However, due to the limitations of the low-cost sensors, the captured meshes are often of low-quality and can only be constructed from 3D data. The richest and most common data source of 2D images is ignored. Xin et al. [4] extend the Mask R-CNN [5] to recover the primitive-shaped objects from a single photograph but human body details are too complex for such methods. The generative-adversarial network is used in [6] to infer 3D dressed human body from a single image but without the underlying body models. Ma et al. [7] generate the human body models in an encoder-decoder way. However, a large amount of pre-aligned human body meshes are required by these methods. Due to the various poses and shapes of human body models, the registration of thousands of meshes is non-trivial, which needs extensive manual intervention and fine-tuning [8], [9]. Moreover, the obtained models of current methods do not present human body shapes in a natural way.

Therefore, in this paper, we attempt to exploit user-generated data that is accessible to general users and achieve aligned, believable and animatable human body models. As shown in Figure 1, noisy human body meshes in arbitrary topology from low-cost scanners and single 2D images taken with smartphones are considered. The high-fidelity human body models are then estimated.

Point clouds While the recent advances in 3D scanning techniques contribute to 3D mesh acquirement, the quality of captured scans tends to be noisy, incomplete and erroneous. A **k**ey **p**oint based approach (**KPhub-PC**) is proposed for **hu**man **b**ody estimation from **p**oint **c**louds based on the human body parametric human body model - SMPL. A sparse set of 57 landmarks are first annotated. A rough initial model is estimated based on the landmarks. In the second stage, instead of directly estimating the optimal shape and pose parameters, we estimate the template model that can best explain the target point clouds and the template is reposed to represent the target.

Images In practice, another common data source is the 2D images we take everyday with smartphones. Therefore, we consider creating high-fidelity human body models based on a single 2D image which is easy to acquire by everyone. To tackle this problem, the **k**ey **p**oint based approach of **hu**man **b**ody estimation from a single **i**mage (**KPhub-I**) is proposed. It optimizes pose and shape of SMPL model so that the projected joints and boundary of the template model are close to the 2D joints and boundary of the human body in the images. Besides, in order to make the estimated model present natural poses in the stable states, a stable pose prior is designed using the Gaussian Mixture model.

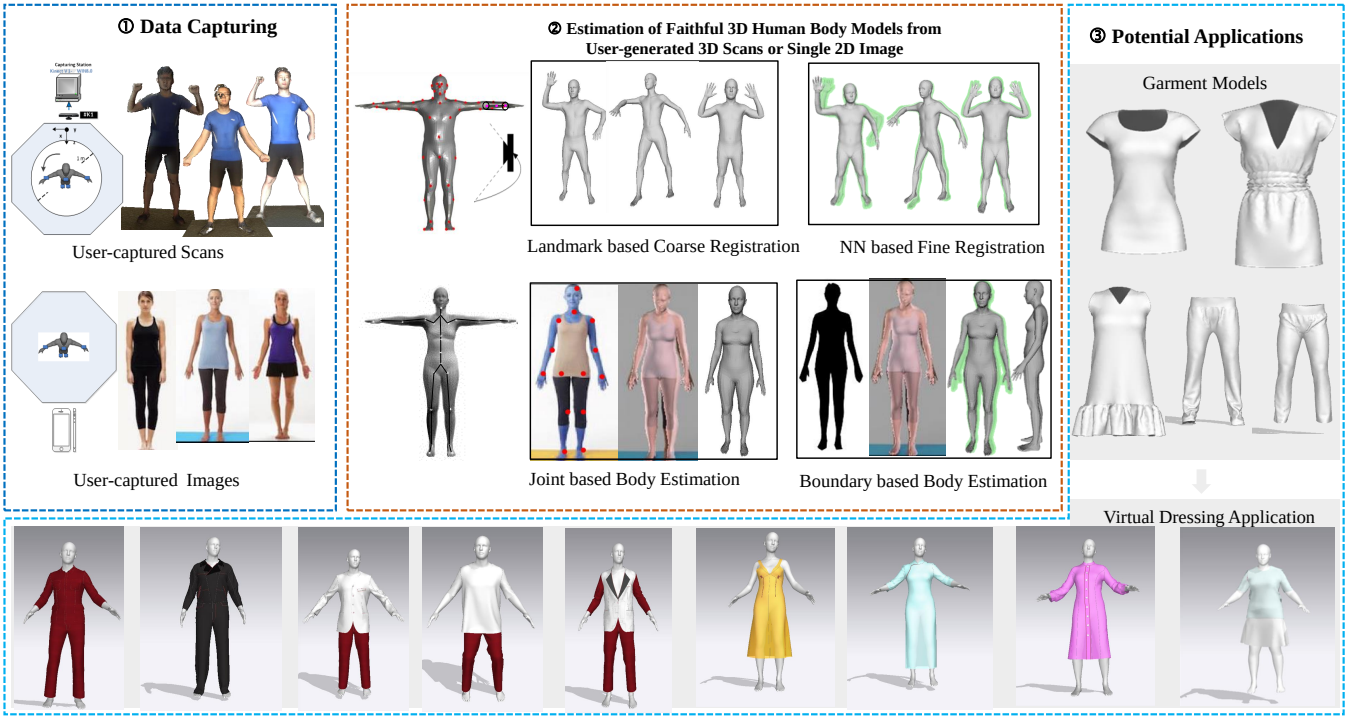


Fig. 1. The workflow of human body modelling for two types of user-generated data sources: the point clouds and single images. The point clouds are captured with a hand-held sensor and the human body images are captured with a smartphone by common users. The template of SMPL model is fitted to the target point clouds or single images to estimate the faithful 3D human body models following a coarse-to-fine process. An example of a potential application is given afterwards. Clothing models are retargeted to the acquired models for personalized virtual dressing applications.

The contributions of this work are listed below.

- **KPhub-PC** is proposed to robustly estimate a template human body model that can best explain the target scans without concerning the topology and can be straightforwardly animatable.

- **KPhub-I** is proposed to estimate the animatable and personalised human avatar from single human body images with 2D joints and boundary information. With a new stable pose prior, our results present more natural poses.

- **SS-hub**, i.e. the Structure Sensor based Human body dataset, is built. To facilitate the development of the research of human body modelling from user-generated scans, a state-of-the-art dataset is collected. SS-hub contains 30 persons in 5 predefined different poses in tight, covering and loose clothes.

- **IM-hub** dataset of images with professional measurements of human bodies is built. To the best of our knowledge, IM-hub is the first human body image dataset with accurate physical measurements, including height, bust width, waist circumference and hip circumference. It contains 50 persons and their measurements. Given the images and the corresponding 3D measurements, IM-hub dataset can be used to quantitatively evaluate the accuracy of the 3D human body shapes which are estimated from images. Both datasets are ready to publish for research purpose (The download link is: https://drive.google.com/drive/folders/1vA-Y1qQTK_OxopqbUJaNQPHfwHB9Lp64?usp=sharing).

II. Related Work

Human body models To simulate human poses and shapes, a large number of human body models have been proposed, ranging from the early-stage stick-like figure and simple geometric primitives based models to current realistic statistic human body models [10], [11], [12], [13]. With the easier acquisition of 3D scans, many systems are proposed to build human body model from 3D scans [14], [15], [16], [17], [8], [18]. The earliest 3D statistical shape model may be [19] where a statistical shape model is trained by performing Principal Component Analysis (PCA) over shape deformation with respect to a template. Allen [20] later improves this work by considering the pose deformation. SCAPE learns two separate models of human pose and human shape and combines them to produce 3D surface models for different people in different poses. BlendSCAPE [21] extends SCAPE by rotating each triangle with a linear blend of the rotation of each part in the skeleton tree. Thus, BlendSCAPE is able to smooth the deformation of the boundaries of two parts. To learn the pose-induced muscle deformation, Chen et al. [18] and Hasler et al. [15] use shared encoding for human body modelling. Stitched puppet [22] builds a part-based human model where each body part is independently translated and rotated rigidly and non-rigidly to fit into the target scans. Then, these parts are stitched together via potential functions. All the above human body models are able to generate realistic human body models. However, their computation is large as their optimizations are non-linear. Based on SCAPE model,

several follow-ups appear, like Stitched puppet [22], Skinned Multi-Person Linear model, short as SMPL [23] and recent SMPL-X model [24]. Joo et al. [16] integrate the modelling of finger and face.

3D human body reconstruction With the above human body models and current development of 3D scanning techniques, many research works reconstruct full-body models by using these existing models as template [9], [25]. Alexiadis [26] present an integrated platform capturing system to reconstruct moving foreground objects, including moving people. Tong et al. [27] use three Kinects to capture the upper, middle and lower parts of a human body. Using the measurements of the first frame, the authors construct a quite rough template using the statistical body shape model in [28]. Then pairwise non-rigidly deformation is performed between successive frames and then deformed back to the first frame. As this system fuses frames from three Kinects, it requires calibration before capturing and several minutes to construct a complete model. In [29], the 3D geometry and appearance of the human body is estimated from a monocular RGB-D sequence of a user moving freely in front of the sensor. It brings the range data into alignment with a parametric 3D body model. Liu et al. [30] automatically build human body models with a template without any human-assigned markers. This work also needs the calibration process in the first place. It requires around 10 seconds to reconstruct a high-resolution human body model. Zhao et al. [1] follow the idea of parametric model fitting for two depth images in the front and the back to reconstruct completed 3D human body models.

As template based human body reconstruction has some limitations when the topology changes greatly or quickly, some researchers propose template free reconstruction approaches [31], [32], [33], [34], [35]. KinectAvtar [36] firstly apply super-resolution algorithm to acquire new super-resolved depth maps with much higher resolution and less noise. Then global rigid and non-rigid alignment steps combine the super-resolved scans into a final model. For each scan, it takes around 14 minutes. 3D Self-portraits [37] implement the scanning of users themselves with a single 3D Kinect by rotating the same pose for a few different views. Then it non-rigidly registers the scans captured in each view into a watertight surface. In [38], objects are reconstructed in high-quality based on a single stream from Kinect but several hours are required to complete. KinectFusion [31] is the first work that implements the reconstruction scenes in real time with commodity scanners. As KinectFusion is proposed under the assumption of static scenes, DynamicFusion [32] implements the real-time reconstruction of dynamic scenes under non-rigid deformation. DynamicFusion inspires a lot of follow-up works, like VolumeDeform [39] which uses sparse RGB feature matching to improve tracking robustness and handle scenes with little geometric variation, allowing for reconstruction of newly emerging parts in real time. DynamicFusion incrementally updates the volumetric representation with new depth input. This reference model confines so that it is hard to reconstruct the cases where quick and dramatic changes happen in topologies and shapes. Therefore, Fusion4D [40] proposes to address the problem of reconstruction with dramatic changes in shapes and topology by taking multiple RGB frames as input. The impressive reconstruction performance makes Fusion4D the supporting technique behind the popular mixed reality application of Holoportation [41]. One of the main problems of the above-mentioned reconstruction systems is that the acquired meshes are static and not directly to be manipulated. Rigging and skinning are required before any pose changing and artifacts are obvious when animation is performed.

Virtual dressing systems More and more human body models can be reconstructed by the above-mentioned template based and template-free reconstruction approaches. Moreover, the advances in e-commercial promote shopping online and the virtual dressing system is indispensable for improving shopping experiences. Many virtual dressing applications appear. Wu et al. [42] propose an image based virtual try-on system. In [43], a laser scanner is used to scan real people to acquire the 3D mesh model and after a series of post-processing like purifying and smoothing, clothes are put on the avatar. In this work, clothes are modeled with clothes scans of real clothes draping on a dressmaker's dummy. The TriMirro [6] systems simulate clothes on a predefined human avatar which can not describe various human body shapes and poses realistically. Fitnect [7] models both human body shape and clothes and allows animation of clothes with pose change. Ye et al. [44] use RGBD data as input to reconstruct a personalized 3D avatar and adapt synthesized clothes. However, the above systems either use a predefined avatar or reconstruct 3D human body models with depth sensors. Consumers must make efforts to go to retail shops where virtual try-on systems are available to try clothes virtually. Moreover, garment is hard to simulated realistically. Current virtual fitting applications do not present the dynamics of apparel. Although ClothCap [45] could be a potential solution for clothes simulation, building models for numerous clothes in the shopping mall is still difficult. The unrealistic simulation of the full human body and clothes limits recommendations.

III. Building High Fidelity Human Body from User-scanned Point Clouds

The availability of 3D scans with low-cost RGBD cameras poses new challenges on building believable human body models from them because of noises, incomplete and obscure parts. Therefore, aiming at building adjustable and high-fidelity human body models from user-scanned point clouds, we propose a robust registration method with the help of a sparse set of landmarks. The overview of the proposed approach of human body modelling from user-generated point clouds is shown in Figure 1. The goal is to fit the SMPL template to the target point clouds to ensure the deformed template can best explain both the shape and pose of the target scans. A sparse set of correspondences is firstly located between human body scans and SMPL template. In the first stage, a rough registration is performed based on those landmarks. In the second registration

[6] http://www.trimirror.com/en/
[7] http://www.fitnect.hu/

stage, nearest neighbour based registration is presented to optimize the templates that can best explain the target scans.

Since the SMPL and target point clouds have different topology and non-isometry, and holes and obscure parts exist on the target scans, the correspondences acquired with current algorithms are error-prone. Therefore, 57 pairs of correspondence are annotated. We follow the landmark placement way in CMU motion capture system [8] and add 16 extra landmarks on the salient points on the whole body, including 9 landmarks on the face and 7 landmarks on the salient points on the main body, like belly button et al. These landmarks not only indicate the poses of the target scan, but they also imply the rough shape of the body. To make full use of the information provided by these sparse landmarks, it is assumed that when the template is close to the target, the correspondences should be completely overlapped from any viewpoint. Therefore, a *projection constraint* is added in the first stage aiming at acquiring closer initialization to the second stage. The second stage of nearest neighbour based registration manages to provide accurate results but is prone to suffer from local minima. To address this problem, an effective *Arm Cylinder Constraint* is proposed. It can keep the human arms' shape robustly, assuring that faithfully human body registration can be acquired.

A. Landmark based Rough Registration

In the first stage, we optimize the shape and pose parameters β and θ of SMPL model to minimize the distance between the SMPL landmarks and the target scans' landmarks. The problem is formulated as:

$$M = \underset{\beta,\theta}{\operatorname{argmin}} E_M(\beta, \theta; L_{ST}), \tag{1}$$

where M is the deformed SMPL model; β and θ are the shape and pose parameters; L_{st} is made up of landmark pairs $\{s_i, t_i\}$ where s_i and t_i are the corresponding landmarks i of SMPL and target scans. We set a camera to see the template as well as the target point clouds from the side view. In this way, we acquire the initial reposed and reshaped SMPL model that is close to target point clouds. Thus, $E_M(\beta, \theta; L_{ST})$ is defined as below.

$$\begin{aligned} E_M(\beta, \theta; L_{st}) =& \lambda_L E_L(\beta, \theta; L_{st}) + \lambda_P E_P(\beta, \theta; K, R, L_{st}) \\ &+ \lambda_\theta E_\theta(\theta) + \lambda_\beta E_\beta(\beta), \end{aligned} \tag{2}$$

where $E_L(\beta, \theta; L_{st})$ is the landmark term; $E_P(\beta, \theta; K, R, L_{st})$ is the projection terms; K and R are the internal parameters of cameras and the camera rotation; E_α and E_β are pose and shape priors respectively, and $\lambda_L, \lambda_P, \lambda_\theta$ and λ_β are weight factors for respective term. These terms are formulated as below.

Data term: Given the landmark correspondences, this term encourages the landmarks of SMPL template to be close to the corresponding landmarks on the target point clouds.

$$E_L(\beta, \theta; L_{st}) = \sum_{landmark_i} (L_{s_i} - L_{t_i}). \tag{3}$$

[8]http://mocap.cs.cmu.edu/info.php

Projected Landmarks Term: Besides minimizing the 3D positions of these landmarks, another projection term is also added: the projected landmarks from the side view and expect that this term is also minimized to enhance the power of constraints from such a sparse landmark sets.

$$E_P(\beta, \theta; K, R, L_{st}) = \lambda_R E_R(\beta, \theta; K, R, L_{st}), \tag{4}$$

in which R means the camera is rotated R from the front view. The energy of the side view is represented as below:

$$E_R(\beta, \theta; K, R_j, L_{st}) = \sum_{landmarki} (\Gamma_{K,R}(L_{s_i}) - \Gamma_{K,R}(L_{t_i})), \tag{5}$$

where $\Gamma(\cdot)$ is the projection function from 3D to 2D that is defined by the camera parameter K and rotation R.

Shape prior term: $E_\beta(\beta)$ is the shape prior learned from the SMPL body shape training set. The shape parameters β are PCA coefficients of a low-dimensional shape space, learned from thousands of registered scans. In this work, we use 10 coefficients.

$$E_\beta(\beta) = \beta^T \Sigma_\beta^{-1} \beta, \tag{6}$$

where Σ_β^{-1} is a diagonal matrix computed with PCA from the shape in the SMPL training set.

B. Nearest Neighbour based Fine Registration

With Eq. 2, the SMPL template is deformed to the target in terms of poses and shapes to some extent. Since the landmarks are so sparse, the deformed template is still plausible. To obtain the accurate template which can believably describe the scanned target, the distance is penalized between the rough registration results acquired from the section III-A and the target scans. To achieve this goal, a second-stage vertex-to-face optimization is proposed based on the deformed SMPL acquired in the initial stage. As holes (missing data) are prevalent in the target point clouds, including big holes on the top and ends of limbs and some holes on the side of the body, if we locate the nearest neighbor points of the template on the target scans, it tends to find error nearest points. Therefore, we minimized the distance between the template point and the plane of its nearest point of the target point clouds. The registration M for each scan can be computed by solving the following optimization problem:

$$M = \underset{T}{\operatorname{argmin}} E_M(T; V_{ST}). \tag{7}$$

To assure high-quality human body meshes can be estimated from target point clouds, we think the following elements should be added into the objective $E_M(T; V_{ST})$. Firstly, the deformed SMPL should be able to explain the target scans, thus, we design the data term E_D to penalize the vertex-to-face distance between the template and the target. To allow for smooth deformation, the transformations of the neighbours should be similar and Laplacian operator $\mathcal{L}$ encodes the neighbour information for the triangular mesh. Therefore, the Laplacian smooth constraints $E_{Laplacian}$ is designed. Besides, the human body moves and interacts non-rigidly so that the modelling of human body suffers from a large number of unknown parameters and the inherent ambiguity, as various

deformations can generate the same shape. This situation can be alleviated by adding the shape and pose priors E_θ and E_β. Another challenge is that in arm parts where good initialization is hard to obtain, local minima problems happen so that the locating of nearest neighbours is error-prone. Inspired by [12] which approximates human arms with simple but effective geometry, we assume that the arm should keep the shape of the cylinder during deformation. Thus, we design the arm term E_A. With the above-mentioned terms, the whole objective $E_M(T; V_{ST})$ is defined as follows.

$$\begin{aligned} E_M(T; V_{ST}) =& \lambda_D E_D + \lambda_L E_{laplacian} \\ &+ \lambda_\theta E_\theta + \lambda_\beta E_\beta + \lambda_A E_A, \end{aligned} \tag{8}$$

where T is the template that to be optimized to best explain the target scans; V_{ST} represents the vertex of the SMPL model and the target scans. Please note that here we do not obtain an optimal shape and pose parameters, but aim at an optimal template model that can explain the target best. The shape prior is defined the same with Eq. 6 and pose prior is defined as [46]. The remaining terms are defined as follows.

1) Vertex-to-face Distance: Data Term measures the distance between vertices of the template and the plane of its nearest point on the target.

$$E_D = \sum_{i=1}^{n} (v_i - t_i) * N_i, \tag{9}$$

where N_i is the normal of the plane determined by the three nearest points of target point i; v_i and t_i are the corresponding points of template and the target respectively.

2) Laplacian Smooth Constraints: We regularize the smooth deformation by adding the Laplacian mesh regularizer [47]. It is defined as:

$$E_{laplacian} = \sum_{i=1}^{n} ||\mathcal{L}(v_i) - \mathcal{L}(v(\beta_{\mathbf{0}})_i)||^2, \tag{10}$$

where $\mathcal{L}$ is the Laplacian operator that describes the local feature for each vertex v_i on the surface; $v(\beta_{\mathbf{0}})_i$ is the vertex i of SMPL model in the initial shape states β_0. The term forces the Laplacian of the optimized mesh to be similar to the Laplacian of the mesh at initialization. The Laplacian is defined in Eq. 11 where d_i is the number of edges attached to vertex v_i; $N_1(i)$ is the neighbour of v_i.

$$\mathcal{L}(v_i) = \frac{1}{d_i} \sum_{j \in N_1(i)} (v_i - v_j) \tag{11}$$

With Laplacian regularization, the local features can be kept during deformation.

3) Approximating Arms with Cylinders: Local minima problem happens often in arm parts where noises, holes and distortions are prevalent and the nearest neighbour is searched by fault. This will produce erroneous and meaningless results. To simulate meaningful arm shapes, we assume that the arm should keep the shape of the cylinder during optimization. To achieve this goal, the arm surface is approximated as a cylinder with an axis and a radius. After performing the first level of rough registration, the arm parts have been fitted by the landmarks which indicate the arm shape. Thus, in the fine level of registration, it is assumed that the distance from the vertices of an arm surface to the axis should keep the same. Two joints are taken - one for the arm and the other for the wrist - to form the axis of the cylinder. As shown in Figure 2, besides the two ending points, the middle point is also located by computing the average position.

With these three key points, the surface points within the radius of the cylinder for each key point are retrieved with kd-trees. The L2 distances from these retrieved surface points to the key points are calculated with the optimized mesh from the rough level of registration.

Arm Cylinder Term: The goal of this term is to keep the shape of human arms for the registration of low-quality scans. The distance is firstly calculated from the searched surface points to the three key points as follows.

$$d = \sum_{i=1}^{3} \sum_{j=1}^{N} ||v_{i,j} - p_i||^2, \tag{12}$$

where N is the number of surface points within the radius of cylinder; $p_i \in \{p_{mid}, p_{wrist}, p_{arm}\}$.

To keep the shape of human arms, the changes of the distance from the axis to the surface points should be minimized. Thus, the arm cylinder regularization term should be defined as:

$$E_A = ||d_{smpl} - d_{rough}||^2, \tag{13}$$

where d_{rough} is the arm-to-key-point distance calculated with Eq. 12 and d_{smpl} is the arm-to-key-point distance of deforming smpl model.

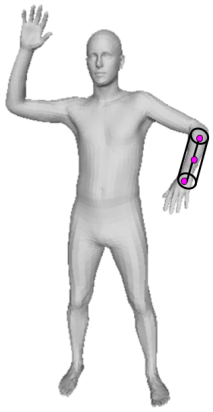


Fig. 2. The arm is approximated by a cylinder.

Implementation Details We optimize the Eq. 1 and Eq. 7 using the $Powell'$s dogleg trust region method, based on the Chumpy [9] and OpenDR [48] auto-differentiation framework. We adopt a stage-based approach to optimize these two objectives. For minimizing Eq. 7, the vertex-to-face correspondences are updated for each stage of optimization. We experimentially choose the weight of each term. λ_θ and λ_β start with a high value and gradually decrease in the later stages to effectively avoid local minima. λ_L and λ_P are set to

[9]https://github.com/mattloper/chumpy

be 200 as the number of landmarks are very small compared with the number of vertex of the whole body. Similarly, λ_A is set as 10. In Eq. 7, λ_D and λ_L are set to be 1 empirically. Optimizing for a mesh takes around 3 minutes on a common desktop machine with 16 GB RAM and 4 cores.

IV. Estimation of 3D Human Body Models from User-captured 2D Images

Although the 3D models can be obtained more easily than before, the scanning systems usually require people to be physically presented where the scanning systems are located. While in the next-generation of digital clothes retailers and fashion industry, consumers expect to remotely generate their realistic 3D human body models with observations of their body shapes and poses using mobile devices that are commonly accessible, like selfies taken by smartphones. However, due to the lack of depth information, estimation of human body shape and pose from a single image is very challenging.

Here, an automatic method, KPhub-I, is proposed to accurately estimate 3D human body meshes from single 2D images and hopefully push a step forward towards online virtual dressing applications. Rather than reconstructing human body avatar based on multiple 2D images from different views which needs careful camera calibration [49] or deep leaning methods [50], [51], [52] which need a large amount of training, as shown in the middle row of Figure 1, the realistic human body avatar is acquired from a single 2D image by fitting the state-of-the-art 3D parametric human body model, SMPL [23] to the image to estimate human body shapes and poses. The SMPLify [46] method is taken as a base method and we go beyond it by exploiting boundary information of images and stable pose space to constrain the deformation of SMPL. With our accurately estimated human body shape and poses, virtual dresses are fit on models to demonstrate the visualization of clothing on the human bodies.

A. Build stable pose prior

In the scenario of virtual try-on, people commonly stand or move slowly in front camera. The pose variance is limited. As CMU dataset covers various human poses presented in daily life and sports for 144 subjects. A general pose prior cannot describe some specific poses accurately. Experiments show that the results of SMPLify present bent knees or stoop for the stable pose of "Stand". In order to provide more accurate pose prior in our case, the stable poses are firstly located from CMU dataset. The stable poses are defined to be those change slightly in a short period of time. For each frame, we calculate the error between its neighboring frames:

$$err = \frac{\sum_{k=-step}^{step} norm(\theta_i - \theta_{i+k})}{2 \times step}. \tag{14}$$

Here, the step is set to be 1 and θ is the pose parameter of the motion in each frame and when err is smaller than $threshold$, the pose is regarded as stable poses. In our experiment, the threshold is set to be 0.001. Mosh [53] is applied to calculate the pose parameters θ for each frame of stable poses, which captures motion and shape from sparse markers provided by CMU mocap data.

Some selected stable pose samples are shown in Figure 3. As can be seen, stable poses support various kinds of poses, including stand, squat, leaning and sitting, which are common poses in a try-on process. With stable poses, we use Gaussian

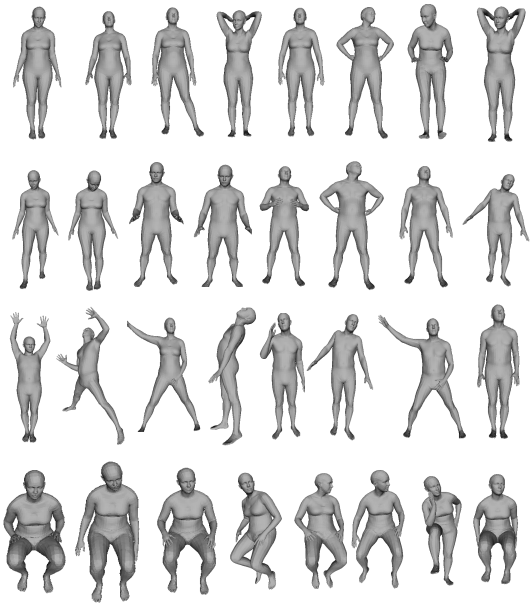


Fig. 3. Sample stable poses.

Mixture Model (GMM) which is a widely-recognised effective model for such a problem to describe the pose prior in our work.

B. Boundary Assisted human body shape and pose estimation

SMPL is taken as our human body representation. Before using boundary information to improve the accuracy of body shape estimation, we utilize joints to estimate the basic figure and poses. Given the estimated 2D joints of the single image J_{est}, the energy function is formulated as:

$$\begin{aligned} E_M(\beta, \theta) = E_J(\beta, \theta; K, J_{est}) + \lambda_\theta E_{S\theta}(\theta) \\ + \lambda_\alpha E_\alpha(\theta) + \lambda_\beta E_\beta(\beta), \end{aligned} \tag{15}$$

where E_J is the data term and $E_{S\theta}$, E_α and E_β are priors.

Data term: this encourages the template M to be close to the target image. For each joint of SMPL template, J_{est_i} we minimize the distance between its projection position and the corresponding image joint. The joint fitting term is formulated as follows:

$$E_J(\beta, \theta; K, J_{est}) = \sum_{jointi} \omega_i(\Gamma_K(R_\theta(J(\beta)_i)) - J_{est,i}), \tag{16}$$

where Γ is the projection function from 3D to 2D that is defined by the camera parameter K; J is the joint estimation function, which returns joint locations; R is the rotation function; ω_i is the confidence we gain from the extraction of joints with DeepCut. Its value depends on the confidence of its estimation. *Shape prior* is the same with Eq. 6.

Stable pose prior term: $E_{S\theta}(\theta)$ and $E_{\alpha}(\theta)$ are pose priors which are learned from precomputed stable poses. Here, $E_{\alpha}(\theta)$ is only performed on the knees to avoid unnatural bending. $E_{S\theta}(\theta)$ can favor probable stable poses over unstable ones. After the stable pose prior is trained in Section IV-A, $E_{S\theta}(\theta)$ is defined as the negative logarithm of a sum. We fit a mixture of Gaussian with 8 components to 30000 poses. As described in [54], a max mixture has much of the same character as a sum mixture and retains a similar expressivity but is well compatible with our optimization framework. Thus, we approximate the sum in the mixture of Gaussian by a max operator:

$$\begin{aligned} E_{S\theta}(\theta) &= -log \sum_{j} (g_j \mathcal{N}(\theta; \mu_{\theta_j} \Sigma_{\theta_j})) \\ &\approx -log(\max_{j} g_j \mathcal{N}(\theta; \mu_{\theta_j} \Sigma_{\theta_j})), \end{aligned} \tag{17}$$

where μ_{θ_j} and Σ_{θ_j} are trained with 30000 stable poses.

Boundary assisted shape estimation After the first stage of estimation with stable pose prior described above, the initial pose and shape have been estimated. The boundary information is very important to enlarge or shrink the model to make the final estimated human body shape similar to the real person. To achieve this goal, the last stage of optimization is defined as:

$$E(\beta, \theta) = E_M(\beta, \theta) + E_b(\beta, \theta; K, U). \tag{18}$$

Boundary term: this encourages the projected boundary of the human body to be close to the image boundary. After we perform two stages of optimization above, the camera position has been estimated, the boundary of SMPL model can be extracted from its projection in the camera. The boundary term is defined as follows:

$$E_b(\beta, \theta; K, U) = \sum_{i}^{N} ||(B_i - U_i(\Pi_K(M(\beta, \theta))))||^2, \tag{19}$$

where B_i is the ith point on the boundary of images, $\Pi(\cdot)$ is the project function and $U_i(\cdot)$ is the corresponding points of B_i on the boundary of projected model. Combined with the eq. 15, we have the complete cost function:

$$\begin{aligned} E(\beta, \theta) &= E_M(\beta, \theta) + E_b(\beta, \theta; K, U) \\ &= E_J(\beta, \theta; K, J_{est}) + E_b(\beta, \theta; K, U) + \lambda_\theta E_{S\theta}(\theta) \\ &+ \lambda_\alpha E_\alpha(\theta) + \lambda_\beta E_\beta(\beta). \end{aligned} \tag{20}$$

During the optimization, the boundary of the projected model is updated for each round of optimization. For further accelerate the convergence of cost function, we first obtain the initial pose and shape from Eq. 15 and start from the optimized model to perform boundary assisted optimization. Similar to the case of point clouds, we empirically choose the values of λ_θ and λ_β by starting with a high value and gradually decrease in the later stages to effectively avoid local minima. The Eq. 15 and Eq. 20 are optimized using $Powell's$ dogleg method, using OpenDR and Chumpy. Optimization for a single image takes around 1 minute on a common desktop machine with 16 GB RAM and 4 cores.

V. Experiments

A. Evaluation of Human Body Modelling from Point Clouds

In this section, the modelling of human body shapes and poses from user-generated point clouds is evaluated. Experiments are conducted on two datasets: a high-quality human body dataset - SCAPE and a low-quality human body scan dataset - structure sensor based human body dataset (SS-Hub). The proposed approach is compared with other three methods: MABR [9]; "KPhub-PC_β_θ" which estimates human body shape and pose parameters in the stage of fine registration and "landmark based SMPL" that is the results of the first stage of KPhub-PC.

1) Datasets: **SCAPE** [14] has 70 registered high-quality human body meshes with a large variance in poses. All these meshes are brought into alignment.

SS-Hub is collected by untrained operators using an economic 3D scanner - structure sensor. In this experiment, only the scans in tight clothes are considered. Each subject has 5 predefined poses which are in line with SCAPE. Thus, the total number of evaluated meshes is 150.

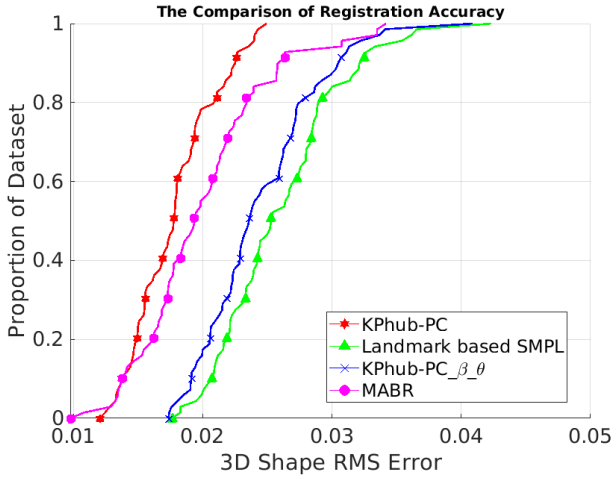


Fig. 4. The comparison of RMS error of Landmark based SMPL, KPhub-PC_β_θ, MABR and the proposed KPhub-PC.

2) Quantitative Evaluation on SCAPE dataset : The superior performance of the proposed KPhub-PC is shown in Figure 4. It is compared with (1) the state-of-the-art MABR which successfully performs human body registration in the case of low-quality Kinect scans, (2) the registration results from the method which only optimizes the SMPL model using landmarks (hereinafter called the 'Landmark based SMPL') and (3) the one that solves for shape and pose parameters in the second optimization stage (hereinafter called the 'KPhub-PC_β_θ'). To compare the estimated human body shape with the observed body shape, we compute the 3D shape root mean square error (RMS Error) with the Eq.21.

$$RMSError = \sqrt{\frac{\sum_{i=1}^{n} (p_i - \hat{p}_i)^2}{n}}, \tag{21}$$

where points p_i and $\hat{p}_i$ are corresponding points of the ground truth and the modeled results, and n is the number of points in the template mesh (SMPL model). As we can see, RMS

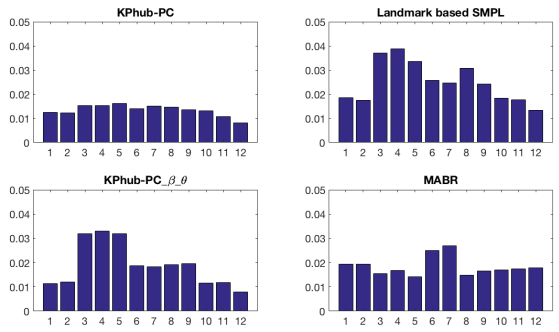


Fig. 5. The comparison of errors for each part of the body. In each subfigure, from left to right, each bar represents the error of: right calf, left calf, right thigh, left thigh, bottom, right forearm, left forearm, torso, breast, right upper arm, left upper arm and head.

Error shows how much the true body shapes vary around the predicted body shapes. It indicates the accuracy of the human body estimation model.

As shown in Figure 4, the accuracy of the proposed approach outperforms the state-of-the-art MABR. Compared with the "Landmark based SMPL", we can see that the initialization is still far from the target and the second nearest neighbour based fine registration in "KPhub-PC" greatly improves the accuracy. Starting from the rough registration results, the "KPhub-PC", "KPhub-PC_β_θ" and "MABR" perform a fine registration. We can see that the accuracy of MABR decreases and this suffers from the local minima due to the bad initialization. Although MABR uses PCA model to improve the robustness against local minimal [9], large variance of poses in SCAPE dataset still poses challenges on such Iterative Closest Point (ICP) based methods. The arm cylinder term effectively preserves the shape of arm, relieving the negative effect from local minima.

Besides, we also compare the errors for each part of the scape model in Figure 5. The body is divided into 12 parts. The average error for each part is calculated. We can see that our overall errors for all parts are more equally-distributed and the values are less than other three methods, which verifies the robustness and accuracy of the proposed methods on all parts of the body. More importantly, the errors of sixth and seventh parts that represent the right and left forearms, are much less than other methods, evidencing the effectiveness of the arm cylinder constraint terms.

3) Qualitative Evaluation on SS-hub Dataset: Since in the evaluation on the raw scans which have no ground truth, we qualitatively show the results of the proposed KPhub-PC, landmark based SMPL and KPhub-PC_β_θ to show the superior performance of our approach. Here only the human body is estimated, thus the hair and the clothes are not presented. In Figure 6, we show the estimated body models as well as the per-vertex error. Please note that as we are lack of the ground truth in the case of raw scans, the per-vertex error is calculated based on the nearest neighbours on the target scans. When the local minima occur, the per-vertex error of the parts that are trapped in local minima is relatively small. Thus, we show both the original estimated models and the per-vertex errors in Column 2-5 and Column 6-9 respectively.

Compared to KPhub-PC_β_θ, the proposed KPhub-PC that obtains an optimal template to best explain the target scans can describe more accurate shape information of the surface. The "Landmark based SMPL" estimates the human body shapes and poses that can best explain the sparse landmarks. Although these sparse landmarks contain some shape information, a lot of details are missing. Therefore, a set of landmarks are not enough to describe body models accurately. As for KPhub-PC without the Arm Cylinder Term, it is obvious to see that it suffers from the local minima that easily occurs in the arm parts. Compared with KPhub-PC without the Arm Cylinder Term, KPhub-PC can produce more faithful arms, improving the robustness of the estimation.

B. Evaluation of Human Body Estimation from a Single Image

1) Evaluation of pose estimation: We quantitatively evaluate the accuracy of 3D pose estimation on the CMU dataset. In order to show the superior performance of the proposed approach, we compare the proposed KPhub-I with three state-of-the-art methods: (1) SMPLify [46] which estimates the body models for single images using SMPL; (2) SMPLify-X [55] which also estimates face, hands, and feet besides body and trains a neural network for pose prior, i.e. VPoser, using a large MoCap dataset; and (3) SPIN [56], a deep learning method to estimate body models from single images using SMPL following an encoder-regressor-optimizer paradigm.

According to the definition of stable poses in Section IV-A, the threshold is set to be 0.001 to find frames of stable poses in CMU mocap dataset. Mosh [53] is firstly performed which captures motion and shape from sparse markers provided by CMU mocap dataset to get pose parameters for each frame which are regarded as ground truth for evaluation. We synthesize the body meshes by giving the ground-truth pose parameters and fixing shape parameters to zeros and then project their joints into 2D with a known camera. The synthesized models are also projected with the same camera on a 2D plane to acquire images. The projected 2D joints will be used as the input of KPhub-I, SMPLify and SMPLify-X. The projected images will be the input of SPIN. We will evaluate the pose-to-pose error. The error is calculated between the ground truth θ_{gt} and the estimated pose θ_{est} parameter based on the formula:

$$e = ||\theta_{gt} - \theta_{est}||^2 \tag{22}$$

In our experiments, 42797 stable poses are found and the 5-fold cross validation is performed for each class to compare the proposed KPhub-I against the state-of-the-art methods SMPLify [46], SMPLify-X and SPIN. The overall average error is presented in Figure 7 which shows the proposed KPhub-I outperforms the state-of-the-art methods. Only 30000 pose data is used in KPhub-I and the most accurate pose estimation is achieved. Other three methods use more general pose priors to favor possible poses over impossible ones. However, given the 2D joints of images, there would be cases that several different 3D joints have the same 2D projection. Thus, the pose prior trained with a wide range of poses cannot prevent such cases.

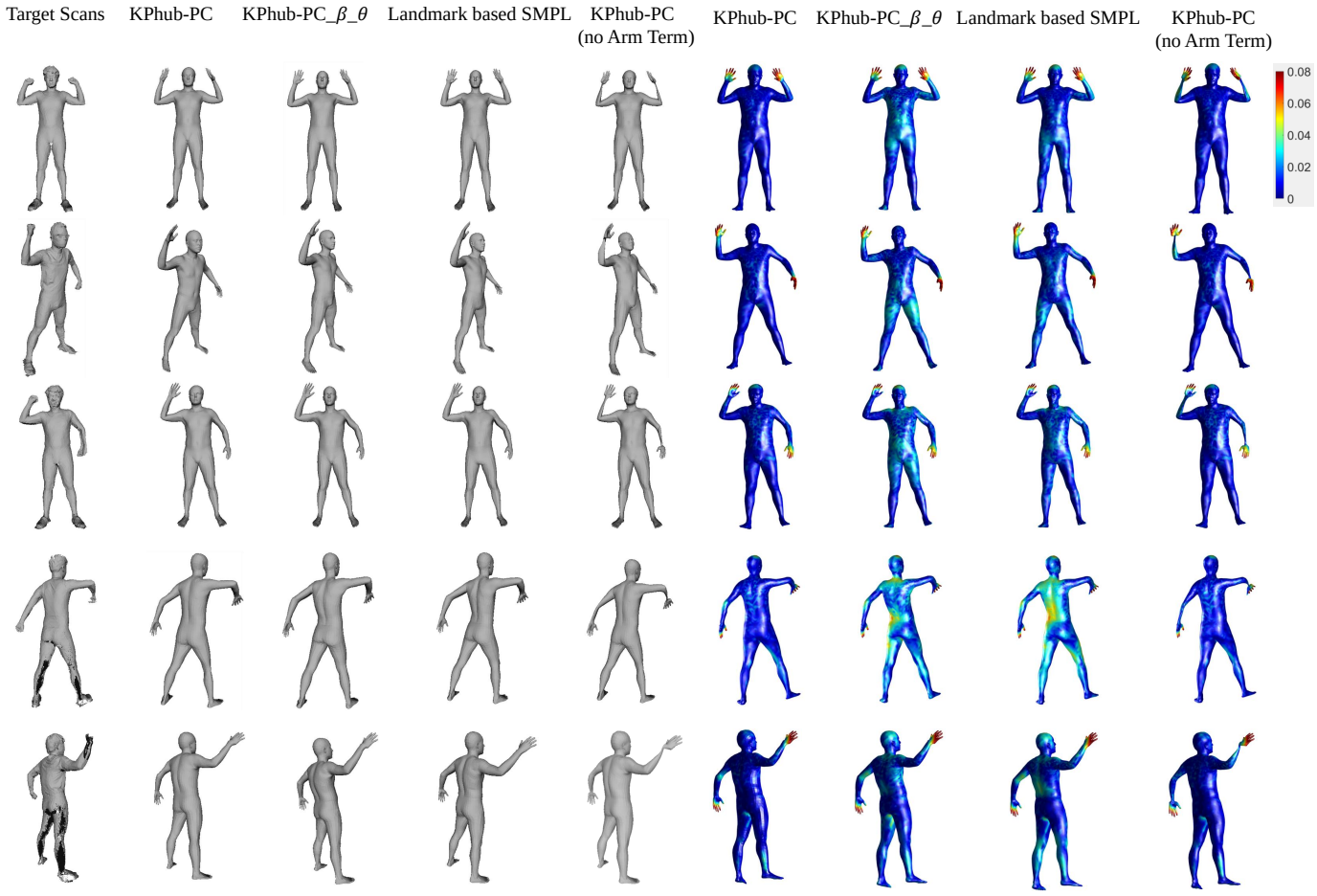


Fig. 6. The qualitative comparison of KPhub-PC, KPhub-PC_β_θ, Landmark based SMPL and KPhub-PC (No Arm Term) on SS-hub Dataset. The estimated models are shown in Column 2-5. The per-vertex errors are compared in Column 6-9.

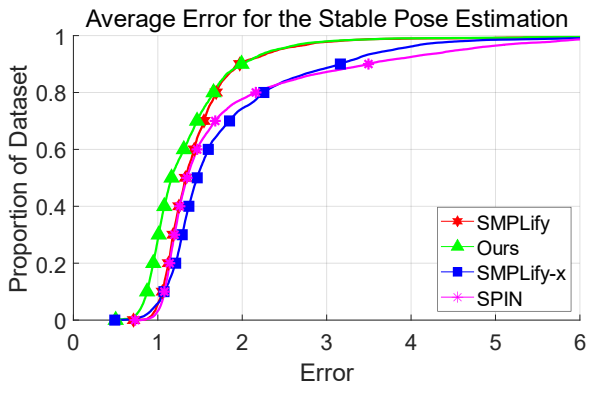


Fig. 7. The comparison of the results of KPhub-I against SMPLify, SMPLify-X and SPIN on stable poses.

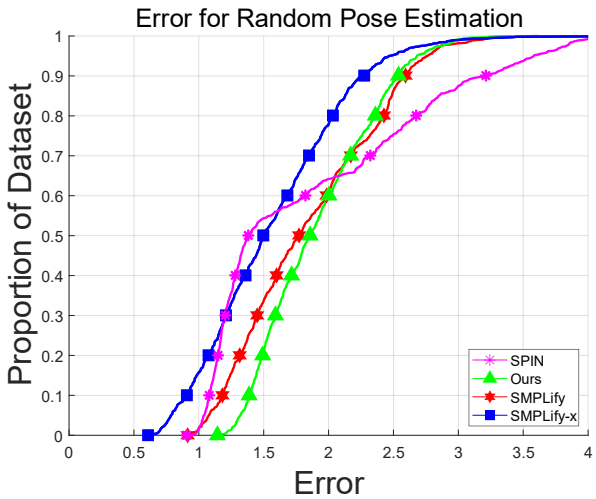


Fig. 8. The comparison of the accuracy of KPhub-I, SMPlify, SMPLify-X and SPIN on random poses.

Besides comparison on stable poses, we also compare the stable pose prior with the general pose prior for random poses. We randomly choose 1000 poses from the CMU dataset and compare pose-to-pose error of methods with stable pose prior and general pose prior on random poses in Figure 8. As we can see, on random poses, SMPLify-X is the most accurate. This is because SMPLify-X trains VPoser as the body pose prior using a variational autoencoder from three publicly available human motion capture datasets including the CMU, Human 3.6M and the PosePrior dataset. The SPIN method is not very stable

in this experiment. However, our method only uses 30000 in CMU dataset and achieves comparable results with SMPLify.

2) Visulization Comparison: More results are visualized in Figure 9 and Figure 10 to show the comparison intuitively. The original images, the overlays of projected model and images, the estimated models and enlarged details of body parts are shown respectively. We compare the proposed KPhub-I with (1) SMPLify, (2) SMPLify-X with SMPL which uses SMPL for body estimation; (3) SMPLify-X with SMPLX which uses a more expressive parametric model of SMPLX for body estimation; and (4) SPIN. Obviously, our method outperforms the state-of-the-arts on stable poses. With the help of the stable pose prior, the estimated results are able to"stand" straight as they are supposed to. Seen from the side, our results are more consistent with what images show while results of other four methods tend to bend knees. In order to compare the projected 2D joints, we also project to 2D plane the whole 3D meshes using the estimated camera and overlap the projected meshes with images. We can see that all joints are projected to the same position on the images. However, four methods generate totally different poses. The comparison verifies that only 2D joints of a single image cannot provide enough constraints on pose estimation.

3) Evaluation on human body measurements: In the above, the accuracy of pose estimation is evaluated. As the ground truth of human body shape is hard to acquire, we compare the shoulder breadth, breast, waist, hip circumferences and height of estimated models with those measured by the professionals. These five measurements are the key elements in the application of virtual try-on. Therefore, to quantitatively compare the accuracy of shape estimation, we collect the dataset of **IM-hub** which contains images of real human and their actual measurements and evaluate the proposed KPhub-I in this dataset.

We compare the results of the proposed approaches, *KPhub-I*, with other two state-of-the-art methods: (1) *SMPLify* [46] and (2) HMR [57], an end-to-end approach to estimate the underlying human body models from images using deep neural network. It is shown that KPhub-I manages to provide more accurate estimation of human body shape and pose overall.

The errors between the ground truth and the measured results for each method are shown in Table I. We calculate the average error of five elements for comparing the overall accuracy. As we can see, for both the male and female, the average error of the proposed KPhub-I is minimal, which verifies the use of boundary information and stable pose prior can improve the accuracy of the estimated models. The deep learning based method, HMR, cannot provide the most accurate results as HMR is trained with a neutral model. If the female and male model is needed, HMR has to be retrained, but only with female/ male images which is not very practical. HMR cannot provide accurate pose estimation so that the height of the estimated model differs a lot from the ground-truth.

4) virtual dressing's look comparision: In order to verify that the accurate estimation of human body shape affects the virtual dressing's look, we first predefine several types of clothes according to the human body measurements and put these predefined clothes on the human body models estimated with SMPLify, SMPLify-X, SPIN and our methods. As shown in Figure 11, we can see the different looks on four estimated body shapes. As the clothes are designed according to the true physical data measured by the professionals, the clothes should fit the body, i.e., not too loose nor too tight. As it can be seen from the stress graph in Column 4, 7, 10 and 13, our method provides the best outfit, whereas the garments on SMPLify and SPIN are too tight and those on SMPLify-X are too loose. Moreover, we also show the front and side views of the dressed models, which verifies that our method presents dressing look more naturally both in terms of shapes and poses. This suggests that our method allows for more accurate shape and pose estimation and the resultant models are more accurate and suitable for virtual dressing.

TABLE I
THE QUANTITATIVE COMPARISON OF THE ACCURACY OF MESH ESTIMATION FOR EACH GENDER. FOR EACH GENDER, THE FIRST ROW IS THE MEASUREMENT ERROR OF THE PROPOSED KPHUB-I; THE SECOND ROW IS THE ERROR OF SMPLIFY; THE LAST ROW IS HMR MEASUREMENT ERROR.

Gender	Methods	Breast	Waist	Hip	Shoulder	Height	Mean
Male	KPhub-I	10.7143	21	8.1429	17.4286	0.6184	**11.5808**
	SMPLify	9.4286	21.2857	10.5714	18.2857	0.1421	11.9427
	HMR	42	14.8571	6.1429	6.1429	5.5424	16.1656
Female	KPhub-I	17.333	29.1667	9	17.6667	0.7241	**14.7782**
	SMPLify	8	32.3333	15.1667	21.1667	0.3079	15.3949
	HMR	46.1667	18.3333	6.1667	21.8333	15.196	21.5383

C. Animation of estimated models

Unlike the results of the traditional animation methods where skeleton has to be embedded and associated with the surface points before animation, our results can be animated straightforwardly with different pose parameters. We show an example of the animated results in Figure 12. The recovered model represents the original scans in the geometry and we further animate it using unseen poses in the dataset (i.e bending and lifting leg). We can see that the animated models present natural geometry.

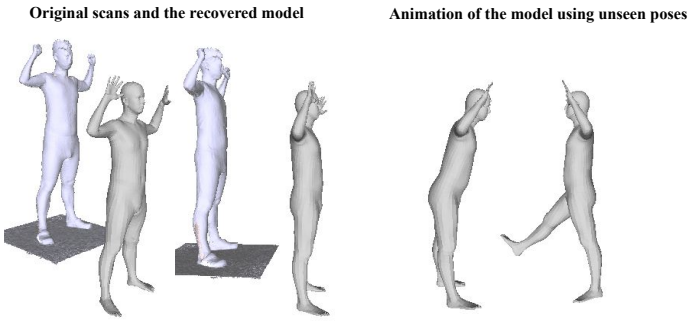


Fig. 12. The example of animated results.

VI. CONCLUSION

In this paper, we present an approach of accurate human body shape and pose estimation for two types of user-generated data: the raw human body scans and single human body images. The estimated human body models can be used

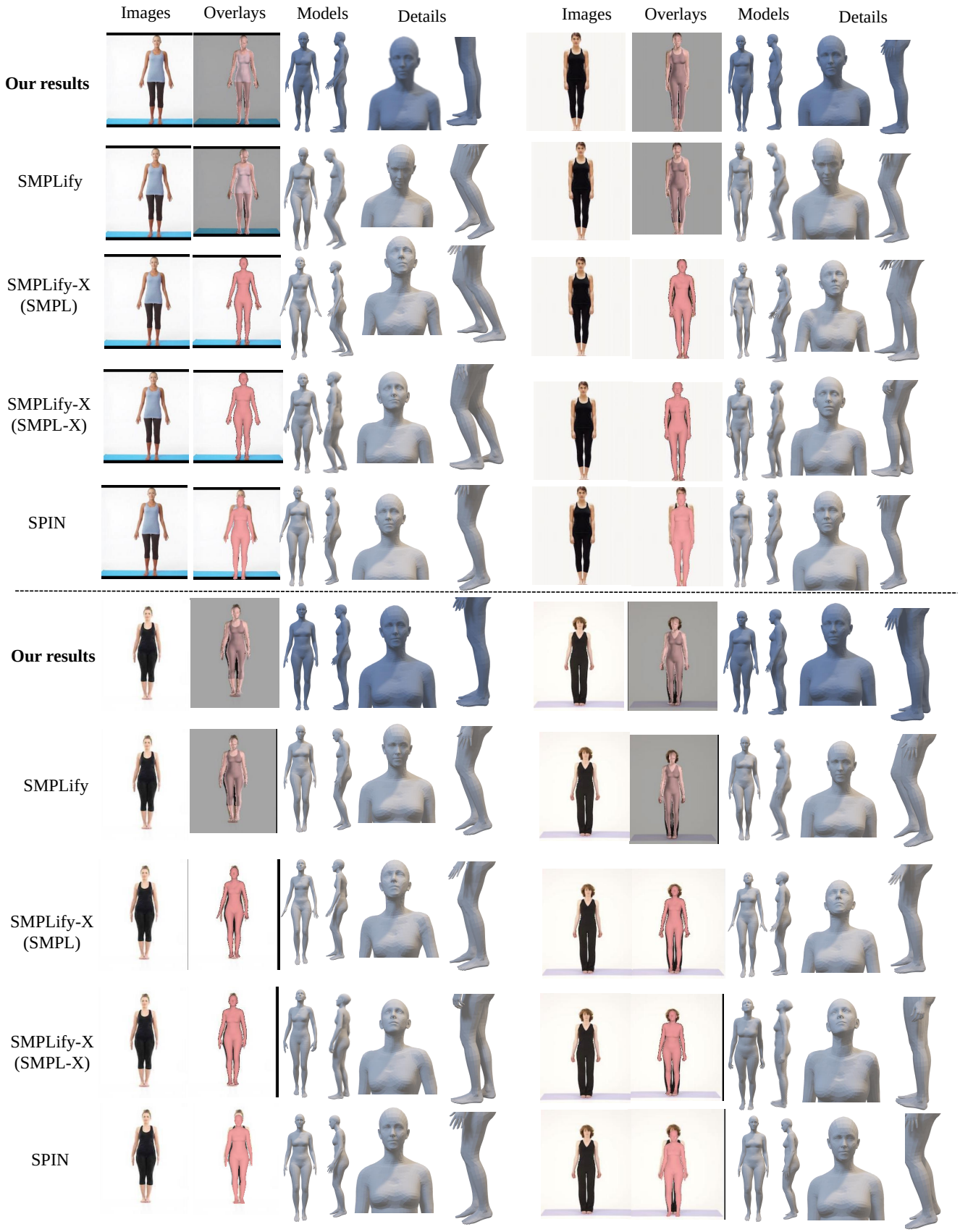


Fig. 9. A qualitative comparison of our results against SMPLify, SMPLify-X with SMPL, SMPLify-X with SMPLX and SPIN results.

Fig. 10. (Continued) The qualitative comparison of our results against SMPLify, SMPLify-X with SMPL, SMPLify-X with SMPLX and SPIN results.

for the scenario of virtual dressing and animation. The robust KPhub-PC is proposed to estimate the accurate naked human body models with the help of a set of sparse landmarks. By introducing the arm cylinder terms, KPhub-PC manages to address the local minima problem effectively. In the case of estimation from a single human body image, we propose to train a stable pose prior and add boundary constraints to this ill-posed problem for a more accurate estimation. Quantitative and qualitative experiments demonstrate that the proposed KPhub-PC and KPhub-I can estimate accurate human body meshes from noisy user-generated data.

As an additional output of this research, two datasets, SS-

Fig. 11. The visualization of predefined clothes model on estimated human body models with Our method, SMPLify, SMPLify-X with SMPLX and SPIN methods.

Hub and IM-Hub, are built. SS-hub contains 450 scans for 30 different subjects. For each subject, 5 predefined poses in tight, covered and loose clothes are captured. IM-hub is the dataset of human body images in front, side and back views and the ground-truth body measurements. We also demonstrate that our works have various interesting applications, including personalised virtual try-on as well as inexpensive bespoke clothing design and manufacture. As future work, it is desirable to develop a parametric clothing model capable of dressing different human body shape and pose automatically.

REFERENCES

[1] T. Zhao, K. N. Ngan, S. Li, and F. Wu, "3-d reconstruction of human body shape from a single commodity depth camera," *IEEE Transactions on Multimedia*, 2018.

[2] Y. Lu, S. Zhao, N. Younes, and J. K. Hahn, "Accurate nonrigid 3d human body surface reconstruction using commodity depth sensors," *Computer animation and virtual worlds*, vol. 29, no. 5, p. e1807, 2018.

[3] F. Mueller, M. Davis, F. Bernard, O. Sotnychenko, M. Verschoor, M. A. Otaduy, D. Casas, and C. Theobalt, "Real-time pose and shape reconstruction of two interacting hands with a single depth camera," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, p. 49, 2019.

[4] C. Xin, Y. Li, X. Luo, T. Shao, J. Yu, K. Zhou, and Y. Zheng, "Autosweep: Recovering 3d editable objects from a single photograph," *IEEE transactions on visualization and computer graphics*, 2018.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[6] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2shape: Detailed full human body geometry from a single image," *arXiv preprint arXiv:1904.08645*, 2019.

[7] Q. Ma, S. Tang, S. Pujades, G. Pons-Moll, A. Ranjan, and M. J. Black, "Dressing 3d humans using a conditional mesh-vae-gan," *arXiv preprint arXiv:1907.13615*, 2019.

[8] F. Bogo, J. Romero, M. Loper, and M. Black, "Faust: Dataset and evaluation for 3d mesh registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3794–3801.

[9] Z. Xu, Q. Zhang, and S. Cheng, "Multilevel active registration for kinect human body scans: from low quality to high quality," *Multimedia Systems*, vol. 24, no. 3, pp. 257–270, 2018.

[10] G. Hinton, "Using relaxation to find a puppet," in *Proceedings of the 2nd Summer Conference on Artificial Intelligence and Simulation of Behaviour*. IOS Press, 1976, pp. 148–157.

[11] S. X. Ju, M. J. Black, and Y. Yacoob, "Cardboard people: A parameterized model of articulated image motion," in *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*. IEEE, 1996, pp. 38–44.

[12] L. Sigal, M. Isard, H. Haussecker, and M. J. Black, "Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation," *International journal of computer vision*, vol. 98, no. 1, pp. 15–48, 2012.

[13] C. Sminchisescu and B. Triggs, "Estimating articulated human motion with covariance scaled sampling," *The International Journal of Robotics Research*, vol. 22, no. 6, pp. 371–391, 2003.

[14] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," in *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3. ACM, 2005, pp. 408–416.

[15] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel, "A statistical model of human pose and body shape," in *Computer Graphics Forum*, vol. 28, no. 2. Wiley Online Library, 2009, pp. 337–346.

[16] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8320–8329.

[17] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele, "Building statistical shape spaces for 3d human modeling," *arXiv preprint arXiv:1503.05860*, 2015.

[18] Y. Chen, Z. Liu, and Z. Zhang, "Tensor-based human body modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 105–112.

[19] B. Allen, B. Curless, and Z. Popović, "The space of human body shapes: reconstruction and parameterization from range scans," in *ACM transactions on graphics (TOG)*, vol. 22, no. 3. ACM, 2003, pp. 587–594.

[20] B. Allen, B. Curless, Z. Popović, and A. Hertzmann, "Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis," in *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*. Eurographics Association, 2006, pp. 147–156.

[21] D. A. Hirshberg, M. Loper, E. Rachlin, and M. J. Black, "Coregistration: Simultaneous alignment and modeling of articulated 3d shape," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 242–255.

[22] S. Zuffi and M. J. Black, "The stitched puppet: A graphical model of 3d human shape and pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3537–3546.

[23] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, p. 248, 2015.

[24] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 975–10 985.

[25] Z. Xu and Q. Zhang, "Region based user-generated human body scan registration," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[26] D. S. Alexiadis, A. Chatzitofis, N. Zioulis, O. Zoidi, G. Louizis, D. Zarpalas, and P. Daras, "An integrated platform for live 3d human reconstruction and motion capturing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 798–813, 2016.

[27] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan, "Scanning 3d full human bodies using kinects," *IEEE transactions on visualization and computer graphics*, vol. 18, no. 4, pp. 643–650, 2012.

[28] N. Hasler, C. Stoll, B. Rosenhahn, T. Thormählen, and H.-P. Seidel, "Estimating body shape of dressed humans," *Computers & Graphics*, vol. 33, no. 3, pp. 211–216, 2009.

[29] F. Bogo, M. J. Black, M. Loper, and J. Romero, "Detailed full-body reconstructions of moving people from monocular rgb-d sequences," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2300–2308.

[30] Z. Liu, J. Huang, S. Bu, J. Han, X. Tang, and X. Li, "Template deformation-based 3-d reconstruction of full human body scans from low-cost depth cameras," *IEEE transactions on cybernetics*, 2016.

[31] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*. IEEE, 2011, pp. 127–136.

[32] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 343–352.

[33] T. Yu12, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu, "Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera," 2017.

[34] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic, "Killingfusion: Non-rigid 3d reconstruction without correspondences," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 3, no. 4, 2017, p. 7.

[35] K. Guo, F. Xu, T. Yu, X. Liu, Q. Dai, and Y. Liu, "Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 3, p. 32, 2017.

[36] Y. Cui, W. Chang, T. Nöll, and D. Stricker, "Kinectavatar: fully automatic body capture using a single kinect," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 133–147.

[37] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev, "3d self-portraits," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, p. 187, 2013.

[38] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi, "3d scanning deformable objects with a single rgbd sensor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 493–501.

[39] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger, "Volumedeform: Real-time volumetric non-rigid reconstruction," in *European Conference on Computer Vision*. Springer, 2016, pp. 362–379.

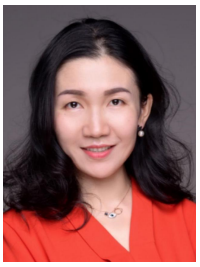
[40] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor *et al.*, "Fusion4d:

Real-time performance capture of challenging scenes," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, p. 114, 2016.
[41] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou *et al.*, "Holoportation: Virtual 3d teleportation in real-time," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 2016, pp. 741–754.
[42] Z. Wu, G. Lin, Q. Tao, and J. Cai, "M2e-try on net: Fashion from model to everyone," in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2019, pp. 293–301.
[43] A. Divivier, R. Trieb, A. Ebert, H. Hagen, C. Gross, A. Fuhrmann, V. Luckas *et al.*, "Virtual try-on topics in realistic, individualized dressing in virtual reality," 2004.
[44] M. Ye, H. Wang, N. Deng, X. Yang, and R. Yang, "Real-time human pose and shape estimation for virtual try-on using a single commodity depth camera." *IEEE transactions on visualization and computer graphics*, vol. 20, no. 4, pp. 550–559, 2014.
[45] G. Pons-Moll, S. Pujades, S. Hu, and M. Black, "Clothcap: Seamless 4d clothing capture and retargeting," *ACM Transactions on Graphics,(Proc. SIGGRAPH)[to appear]*, vol. 1, 2017.
[46] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *European Conference on Computer Vision*. Springer, 2016, pp. 561–578.
[47] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel, "Laplacian surface editing," in *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*. ACM, 2004, pp. 175–184.
[48] M. M. Loper and M. J. Black, "Opendr: An approximate differentiable renderer," in *European Conference on Computer Vision*. Springer, 2014, pp. 154–169.
[49] H. Jin, S. Soatto, and A. J. Yezzi, "Multi-view stereo reconstruction of dense shape and complex appearance," *International Journal of Computer Vision*, vol. 63, no. 3, pp. 175–189, 2005.
[50] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction," in *European Conference on Computer Vision*. Springer, 2016, pp. 628–644.
[51] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 945–953.
[52] Z. Lun, M. Gadelha, E. Kalogerakis, S. Maji, and R. Wang, "3d shape reconstruction from sketches via multi-view convolutional networks," *arXiv preprint arXiv:1707.06375*, 2017.
[53] M. Loper, N. Mahmood, and M. J. Black, "Mosh: Motion and shape capture from sparse markers," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, p. 220, 2014.
[54] E. Olson and P. Agarwal, "Inference on networks of mixtures for robust robot mapping," *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 826–840, 2013.
[55] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. Osman, D. Tzionas, and M. Black, "Expressive body capture: 3d hands, face, and body from a single image," 06 2019, pp. 10 967–10 977.
[56] N. Kolotouros, G. Pavlakos, M. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *Proceedings of the IEEE international conference on computer vision*, 09 2019, pp. 2252–2261.
[57] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7122–7131.

Zongyi Xu received the Ph.D.degree from Queen Mary University of London (QMUL) in January 2019, the Master's Degree in Computer Technology and the Bachelor's Degree in Information Security from the University of Electronic Science and Technology of China. She is currently a lecturer at Chongqing University of Posts and Telecommunications. Her research interests are Multimedia, Computer Vision, 3D Human Body Modelling, Nonrigid Registration.

Wei Chang is an associate professor in the school of Art and Design, Beijing Institute of Fashion and Technology. His research interests are low-tech animation; Creative practice and theory of Visual art; The Art Creation and Application of the Integration of Animation and Multimedia Technology in theatre, fashion show, architecture and digital Space; The application of digital creation in the fashion industry.

Yindi Zhu is a fashion designer and an external teacher in Zhijiang College of Zhejiang University of Technology. She graduated from Beijing Institute of Fashion Technology as a Master. Her research interests are virtual garment design; creative draping and innovative design of traditional clothing.



Le Dong received the Ph.D. degree in electronic engineering and computer science from Queen Mary University of London in 2009. She is currently a Professor with the University of Electronic Science and Technology of China. She has authored dozens of papers in international journals and conferences including several top journals and high level international conferences, such as TIP, ACM MM, TMM, PR, TCSVT, and ICPR.

Huiyu Zhou received the B.E. degree in radio technology from the Huazhong University of Science and Technology, China, the M.Sc. degree in biomedical engineering from the University of Dundee, U.K., and the D.Phil. degree in computer vision from Heriot-Watt University, Edinburgh, U.K. He is currently a Professor with the School of Informatics, University of Leicester, United Kingdom. He has published widely in the field.

Qianni Zhang received the Ph.D.degree from Queen Mary University of London, U.K., in 2007. She is currently a senior lecturer at the School of Electronic Engineering and Computer Science, Queen Mary University of London. Her research interests include medical image analysis and understanding; deep learning networks for immunohistochemical data classification, 3D human modelling and animation, augmented reality for surgery planning and guidance.